high resolution MS data directly onto pathways, (ii) cross-integration of genomic and proteomic data and (iii) metabolite identity verification via data-dependent MS/MS analysis, either separately or as part of the autonomous workflow[5].

Our multi-omic analysis tool uses embedded BioCyc[4] and Uniprot[6] databases to map user-uploaded gene and protein data onto the predicted metabolic pathways (**Supplementary Fig. 1**). Results can be viewed in table form or using the interactive Pathway Cloud plot (**Fig. 1**). Dysregulated pathways with greater percent overlap and statistical significance appear in the upper right of the cloud plot. Graph features can be clicked to view more information on overlapping gene, protein and metabolite data, with links to BioCyc, KEGG and METLIN. Important features can be readily identified, helping to decipher underlying biological mechanisms. Details on the pathway analysis and integrated omics workflow can be found in the **Supplementary Methods**. Data sharing is possible between collaborators and the public, and we encourage users to share their data in the XCMS Online community.

To demonstrate metabolic pathway analysis and multi-omic integration, we describe representative sample sets in the **Supplementary Note**, including metabolic pathway analysis using progenitor cell proliferation data and a bacterially induced corrosion study (**Supplementary Fig. 2**); proteomic integration with an aging study (**Supplementary Fig. 3**); transcriptomic and proteomic integration using a human colon cancer study (**Supplementary Fig. 4** and **Supplementary Table 1**); a nitrate stress response study in sulfate-reducing bacteria (**Supplementary Fig. 5**) and a media stress response study in *Escherichia coli* (**Supplementary Fig. 6** and **Supplementary Table 2**); and a cohort of 1,600 diabetes plasma samples (**Supplementary Fig. 7**), which helps illustrate the scalability of the cloud-based XCMS Online.

Other notable tools providing pathway analysis and multi-omic integration include Galaxy-M[7], Open MS from KNIME[8] and MetaboAnalyst[9]. However, many of these tools still require separate preprocessing of tandem liquid chromatography—mass spectrometry data and are not fully integrated into a single program. Our workflow automatically maps metabolomic data directly onto pathways and integrates transcriptomics and proteomics for systems-wide interpretation in one cohesive platform. Additionally, metabolic network mapping is available based on the predictive activity network algorithm[3] for analysis of metabolomic data only, with multi-omics networking in development. In the future, we will incorporate unique metabolic pathways and networks from other sources to provide more comprehensive biological resources.

Data availability. To assist users with the workflow, we have provided a sample data set entitled "Ecoli_glucose-vs-adenosine" (Job ID #1133019) that can be found on XCMS Online under XCMS Public (https://xcmsonline.scripps.edu/landing_page.php?pgcontent=listPublicShares), as well as two instructional videos available on the XCMS Institute website (https://xcmsonline.scripps.edu/landing_page.php?pgcontent=institute) under the Omics tab and by clicking Integrated Omics or Pathway Cloud Plot.

ACKNOWLEDGMENTS

The authors thank J. Nazroo, G. Tampubolon, N. Pendleton and F.C.W. Wu from the University of Manchester for constructive discussions alongside Medical Research Council grant MRC G1001375/1 (R.G.) for generous funding. The authors thank the following for funding assistance: Ecosystems and Networks Integrated with Genes and Molecular Assemblies (ENIGMA), a Scientific Focus Area Program at Lawrence Berkeley National Laboratory for the US Department of Energy, Office of Science, Office of Biological and Environmental Research under contract number DE-AC02-05CH11231 (G.S.); and National Institutes of Health grants R01 GMH4368 (G.S.) and P01 A1043376-02S1 (G.S.).

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

AUTHOR CONTRIBUTIONS

T.H., E.M.F., D.R., H.P.B., A.A., B.H., T.I., M.W.W.A., P.D.R., L.J.N., M.W.F. and G.S. contributed to multi-omic platform design and development; T.H., E.M.F., C.H.J., M.F., G.K., M.P.T., L.L.L., F.L.P. E.L.M., J.D.W., N.J.W.R. and R.G. contributed to data collection and analysis. T.H. and E.M.F. share first authorship. C.H.J., J.I., T.I., and G.S. also contributed to manuscript writing.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Tao Huan[1,13], Erica M Forsberg[1,13], Duane Rinehart[1], Caroline H Johnson[1,2], Julijana Ivanisevic[3], H Paul Benton[1], Mingliang Fang[1,4], Aries Aisporna[1], Brian Hilmers[1], Farris L Poole[5], Michael P Thorgersen[5], Michael W W Adams[5], Gregory Krantz[6], Matthew W Fields[6], Paul D Robbins[7], Laura J Niedernhofer[7], Trey Ideker[8], Erica L Majumder[9], Judy D Wall[9], Nicholas J W Rattray[2,10], Royston Goodacre[10], Luke L Lairson[11] & Gary Siuzdak[1,11,12]

[1]Scripps Center for Metabolomics, The Scripps Research Institute, La Jolla, California, USA. [2]Yale School of Public Health, Yale University, New Haven, Connecticut, USA. [3]Metabolomics Platform, Faculty of Biology and Medicine, University of Lausanne, Lausanne, Switzerland. [4]School of Civil and Environmental Engineering, Nanyang Technological University, Singapore. [5]Department of Biochemistry and Molecular Biology, University of Georgia, Athens, Georgia, USA. [6]Department of Microbiology and Immunology and Center for Biofilm Engineering, Montana State University, Montana State University, Bozeman, Montana, USA. [7]Departments of Metabolism and Aging, The Scripps Research Institute-Florida, Jupiter, Florida, USA. [8]Department of Medicine, University of California San Diego, La Jolla, California, USA. [9]Department of Biochemistry, University of Missouri, Columbia, Missouri, USA. [10]Manchester Institute of Biotechnology, School of Chemistry, The University of Manchester, Manchester, UK. [11]Department of Chemistry, The Scripps Research Institute, La Jolla, California, USA. [12]Departments of Molecular and Computational Biology, The Scripps Research Institute, La Jolla, California, USA. [13]These authors contributed equally to this work.
e-mail: siuzdak@scripps.edu

1. Gowda, H. *et al. Anal. Chem.* **86**, 6931–6939 (2014).
2. Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R. & Siuzdak, G. *Anal. Chem.* **78**, 779–787 (2006).
3. Li, S.Z. *et al. PLoS Comput. Biol.* **9**, 7 (2013).
4. Caspi, R. *et al. Nucleic Acids Res.* **42**, D459–D471 (2014).
5. Benton, H.P. *et al. Anal. Chem.* **87**, 884–891 (2015).
6. The UniProt Consortium. *Nucleic Acids Res.* **43**, D204–D212 (2015).
7. Davidson, R.L., Weber, R.J.M., Liu, H.Y., Sharma-Oates, A. & Viant, M.R. *Gigascience* **5**, 10 (2016).
8. Aiche, S. *et al. Proteomics* **15**, 1443–1447 (2015).
9. Xia, J., Sinelnikov, I.V., Han, B. & Wishart, D.S. *Nucleic Acids Res.* **43**, W251–W257 (2015).

Addressing reproducibility in single-laboratory phenotyping experiments

To the Editor: Phenotyping genetically engineered mouse lines has become a central strategy for discovering mammalian gene function. The International Mouse Phenotyping Consortium (IMPC) coordinates a large-scale community effort for phenotyping thousands of mutant lines[1], making data accessible in public databases[2] and distributing novel mutant lines as animal models of human diseases. The utility of any findings, however, critically depends on whether they can be replicated in other laboratories. This 'megascience' project is but one example of the general concern regarding replicability[3]. Here we introduce a statistical approach and implementation (https://stat.cs.tau.ac.il/gxl_app/) that can be used to estimate the interlab replicability of new results in a single laboratory.

An influential multilaboratory phenotyping study[4] concluded that "experiments characterizing mutants may yield results that are idiosyncratic to a particular laboratory" on account of significant genotype-by-laboratory interaction ($G \times L$) in several phenotypes. However, we proposed[5] a more appropriate statistical model ascribing random effect to each laboratory and its interaction with genotype. This 'random lab model' (RLM) considers the laboratories in the study as a sample representing all potential phenotyping laboratories. It therefore adds the interaction 'noise' $\sigma^2_{G \times L}$ to the individual animal noise, generating an adjusted yardstick against which genotype differences are tested. Consequently, RLM raises the benchmark for finding a significant genotype effect, trades some statistical power for ensuring replicability, and widens the confidence interval of the estimated effect size (**Supplementary Fig. 1**).

In practice, however, almost all preclinical experiments are single-lab studies. Suppose that a researcher phenotypes an important animal model and makes a discovery that the difference between the phenotypes of mutants and wild-type controls is large and statistically significant. How would researchers in other labs know whether to use this mutant and expect to replicate the effect? The RLM also implies that in single-lab experiments the correct yardstick against which the genotype effect is tested should include $\sigma^2_{G \times L}$ in addition to the commonly used within-lab variability. We term this a '$G \times L$ adjustment' (**Supplementary Methods**; implications demonstrated in **Fig. 1a–c**) and validate it by analyzing eight data sets from published multilab mouse phenotyping studies and databases (**Supplementary Table 1** and **Supplementary Note 1**). These data sets include standard physiological, anatomical, and behavioral phenotypes measured in inbred strains and mutant lines. They offer the opportunity to assess the replicability of single-lab results against the multilab RLM conclusions regarding replicable genotypic difference.

From each laboratory's point of view, we compare the $G \times L$ adjustment method with the standard method of analysis, a two-tailed t-test at 0.05 significance level using within-lab variability. Cases in which the RLM analysis did not indicate a replicated genotype effect enabled us to quantify false discoveries (type-I errors; note that we term a nonreplicable discovery 'false' simply because it proved idiosyncratic to the laboratory discovering it). Over all data sets, the average type-I error rate of the standard method ranged between 19.3% and 41% (**Fig. 1d**). This can be viewed as an estimate of the replicability situation in the field of mouse phenotyping, assuming the high standardization level in these data sets. $G \times L$ adjustment reduced this error rate to the vicinity of the chosen 0.05, ranging from 3.3% to 9%, at the cost of reducing power by 8–30% (**Supplementary Table 1**). Potential biases in the above estimates were addressed using a simulation study (**Supplementary Note 2**).

For brevity, we present $G \times L$ adjustment by way of statistical significance and type-I errors, but the same adjusted standard error should be used to construct replicable confidence intervals. Comparison of multiple phenotypes requires that FDR be applied to the $G \times L$ adjustment P values (as in ref. 6). Similarly, the error rate of 'hits' reported by IMPC is lower than those in **Figure 1d**, because the IMPC imposes a considerably more conservative significance threshold.

Whereas here, we $G \times L$-adjust using $\sigma^2_{G \times L}$ estimated from the multilab analysis, general use will employ $\sigma^2_{G \times L}$ from previous multilab studies, with similar phenotypes but possibly other genotypes, treating $\sigma^2_{G \times L}$ as a property of the phenotype rather than of the genotype. This procedure is practical; phenotyping only a few genotypes across several laboratories enables $\sigma^2_{G \times L}$ estimation and adjustment of other genotypes in other laboratories. No highly coordinated collaboration is required, as these results can merely be posted in a combined database for the benefit of the community. We provide a prototype web server for performing $G \times L$ adjustment (https://stat.cs.tau.ac.il/gxl_app/). By entering phenotypic results and testing conditions, users receive $G \times$

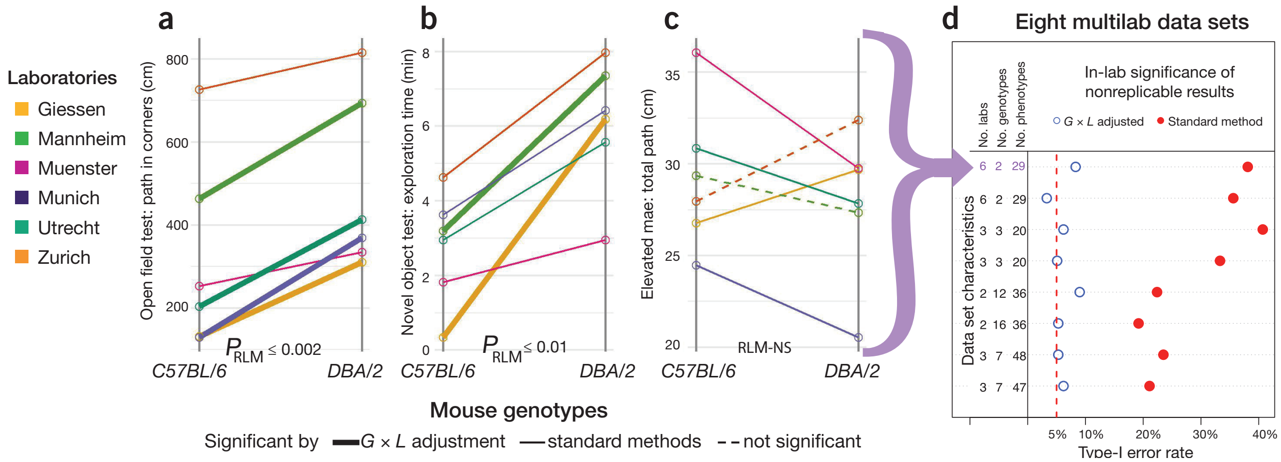

Figure 1 | Adjusting for genotype-by-laboratory interaction ($G \times L$). (**a**–**c**) Comparison of two mouse genotypes for three phenotypes across six laboratories from data set 1 (**Supplementary Table 1**). Each line connects genotype means within the same laboratory, so the slope of each line reflects the difference in these means. Within-lab significances (coded by line type) are all two sided at 0.05. (**a**) Small $G \times L$ effect (similar slopes) and significant genotype effect according to the Random Lab Model (RLM). (**b**) Moderate $G \times L$ effect (more variation among labs), but genotype effect appears fairly replicable and is significant according to the RLM. (**c**) Substantial $G \times L$ effect and no significant genotype effect according to RLM. Standard single-lab analysis would report significant genotype effect for Giessen that is opposite in direction to significant effects for Mannheim, Muenster, and Munich. (**d**) $G \times L$ adjustment decreases percentage of nonreplicable discoveries (type-I error rate).

L-adjusted P values and confidence intervals from available relevant estimates, and they are encouraged to post their results in order to further enrich the database (see **Supplementary Fig. 2**).

A similar approach 'simulates' $\sigma^2_{G \times L}$ by systematically 'heterogenizing' housing and testing conditions within single-lab studies[7]. As indicated by the consistently lower type-I error rate in the heterogenized data set (data sets 2 versus 1 in **Supplementary Table 1**), this may be a worthwhile effort, although the simple form of heterogenization used did not capture all of the estimated $\sigma^2_{G \times L}$.

The concern about replicability of phenotyping results may be regarded as an example of the general concern about reproducibility in science, which has been attributed to issues such as the file drawer effect, publication bias, financial and publicity incentives, etc.[4]. While these are all relevant problems, substantial statistical issues have yet to be addressed, such as testing with the relevant variability as discussed here. The $G \times L$-adjusted P value and confidence interval indicate the prospects of replicating the result in additional laboratories. Reporting these values side by side with the usual P value and confidence interval will promote replicability of preclinical research.

Data availability statement. All data and analysis are publically available; see "S.3 Data and Code Availability" in the Supplementary Information.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This work is supported by European Research Council grants FP7/2007-2013 ERC agreement no[294519]-PSARPS (Y.B., N.K., I.G., I.J., T.S., and S.Y.) and REFINE (H.W.). We thank the International Mouse Phenotyping Consortium (IMPC) and their Data Coordination Centre for the provision of phenotyping data sets.

AUTHOR CONTRIBUTIONS

N.K., I.G., and Y.B. conceived the project in cooperation with all other authors. All authors contributed and/or extracted previously published data. I.J., H.M., T.S., S.Y., and Y.B. performed statistical analyses and programmed software. N.K., I.J., and Y.B. drafted the paper, to which I.G., H.M., T.S., H.W., and S.Y. also contributed.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Neri Kafkafi[1], Ilan Golani[2], Iman Jaljuli[1], Hugh Morgan[3], Tal Sarig[4], Hanno Würbel[5], Shay Yaacoby[1] & Yoav Benjamini[1,6]

[1]Department of Statistics and O.R., School of Mathematical Sciences, Tel Aviv University, Tel Aviv, Israel. [2]Department of Zoology, Tel Aviv University, Tel Aviv, Israel. [3]MRC Harwell, Mammalian Genetics Unit, Oxfordshire, UK. [4]Department of Statistics and Data Sciences, Yale University, New Haven, Connecticut, USA. [5]Division of Animal Welfare, Vetsuisse Faculty, University of Bern, Bern, Switzerland. [6]The Sagol School of Neuroscience and The Edmond J. Safra Center for Bioinformatics, Tel Aviv University, Tel Aviv, Israel.
e-mail: ybenja@post.tau.ac.il.

1. de Angelis, M.H. *et al. Nat. Genet.* **47**, 969–978 (2015).
2. Koscielny, G. *et al. Nucleic Acids Res.* **42** D1, D802–D809 (2014).
3. Collins, F.S. & Tabak, L.A. *Nature* **505**, 612–613 (2014).
4. Crabbe, J.C., Wahlsten, D. & Dudek, B.C. *Science* **284**, 1670–1672 (1999).
5. Kafkafi, N., Benjamini, Y., Sakov, A., Elmer, G.I. & Golani, I. *Proc. Natl. Acad. Sci. USA* **102**, 4619–4624 (2005).
6. Richter, S.H., Garner, J.P. & Würbel, H. *Nat. Methods* **6**, 257–261 (2009).
7. Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N. & Golani, I. *Behav. Brain Res.* **125**, 279–284 (2001).